The Probability Distribution Features of the Computer Network Traffic Components

A.A. Makarov, G.I. Simonova,

N.L. Kovba

Mechanics institute, MSU

The main reasons to investigate the traffic components distribution

- the network hardware and software regulation and testing;
- the long-time planing and network modernization;
- the quality of service;
- the intrusion detection

Some our results and publications of these studies are presented at:

http://models.msu.ru

The object of our investigations: Russian scientific and educational networks - RBNet

RBNet consolidate about 100 regional networks

All illustrations in this report concern the input of the magistral chanel RBNet- Teleglobe.

The general parameters of the input traffic

- 1. IP – the number of packets;
- 2. IB – the volume of transmitted information in bytes;
- 3. IF – the number of the flows;
- 4. IT – the total time of the flows;
- 5. The average size of the packets;
- 6. Other parameters.

The various sections of the traffic

- 1. By the type of the network protocol (http, ftp et al.);
- 2. By the sub-networks;
- 3. By the individual IP addresses;
- 4. Etc.

The common used types of traffic aggregation

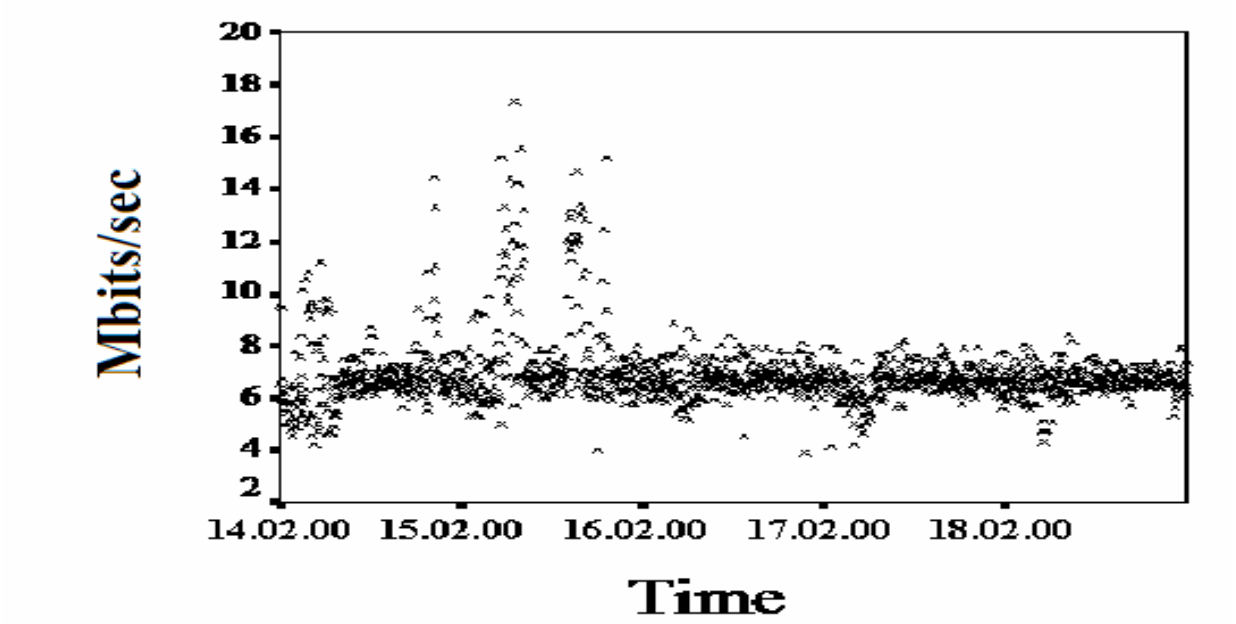
- non aggregated data
- five minite aggregated data
- hour-aggregated data
- day aggregated data

Different aggregation levels for various network problems is used.

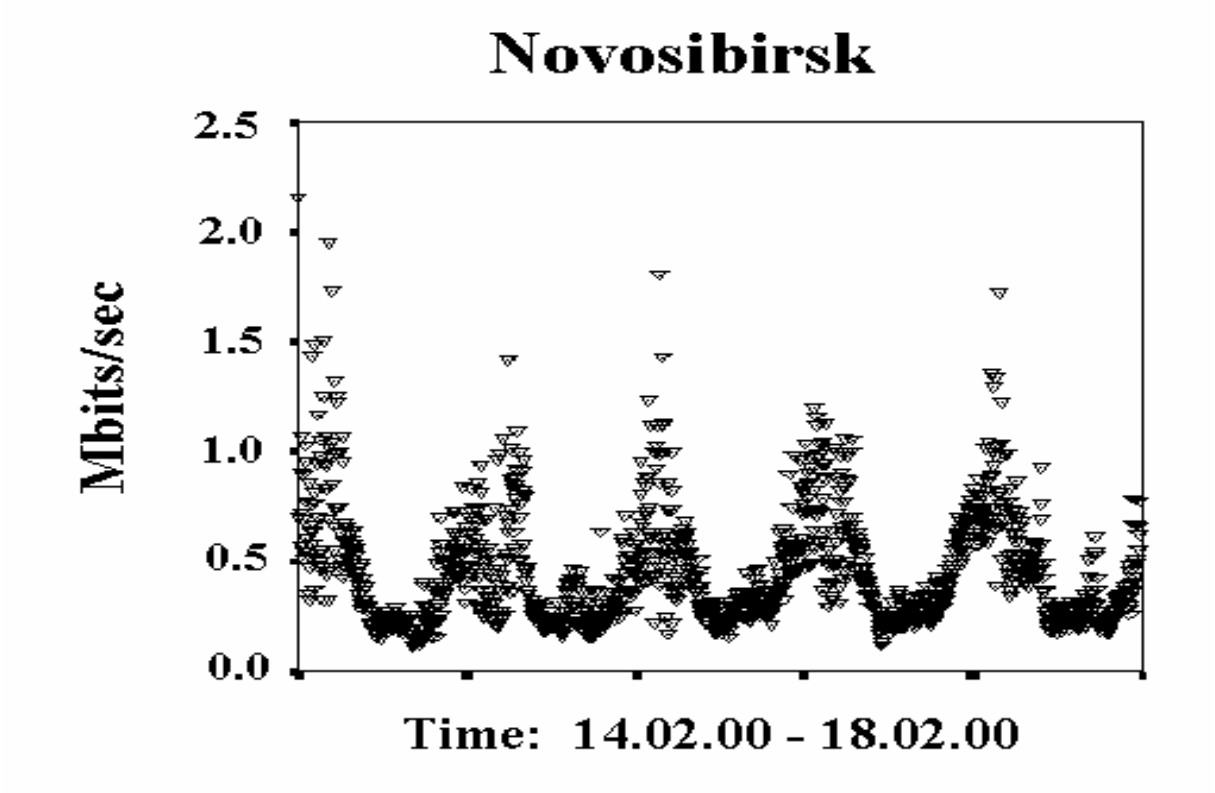
The typical features of the distributions of the five-minute aggregated traffic components

- the nearly observations are independent or weakly dependent due to range of aggregation
- the mean, variance and distribution form are time dependent, but the local stationarity periods exist;
- the general part of data is localized in the within narrow interval relative to the general range of the data;
- the absence of appropriate parametric family for the description of the distribution law;
- asymmetry;
- the small part of extreme, “unusual” observations present in data.

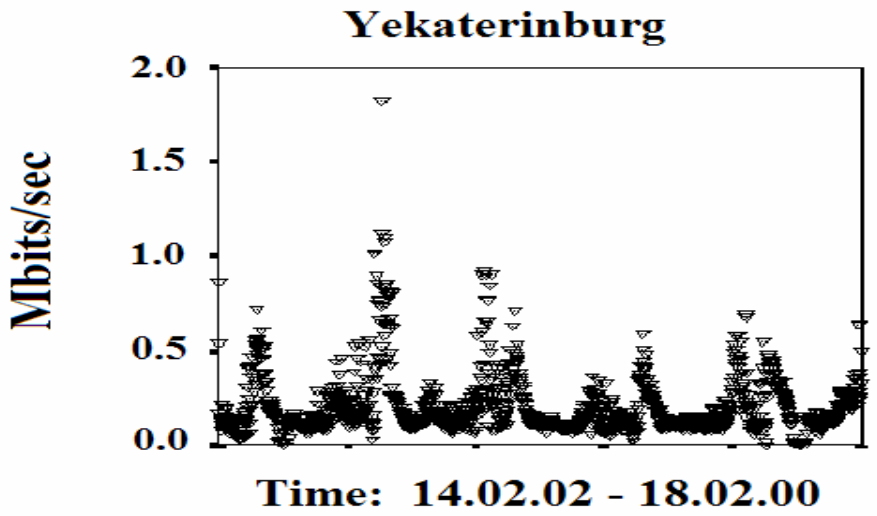
The typical traffic of the RBNet-Teleglobe input till 5 days (5 min aggregation)



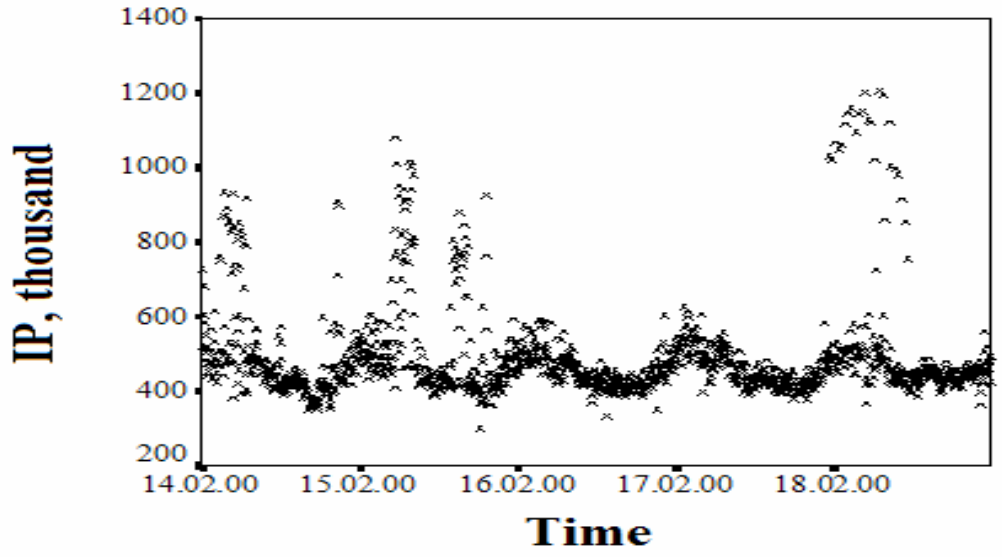
The typical traffic of Novosibirsk at the RBNet-Teleglobe input till 5 days (5 min aggr.)



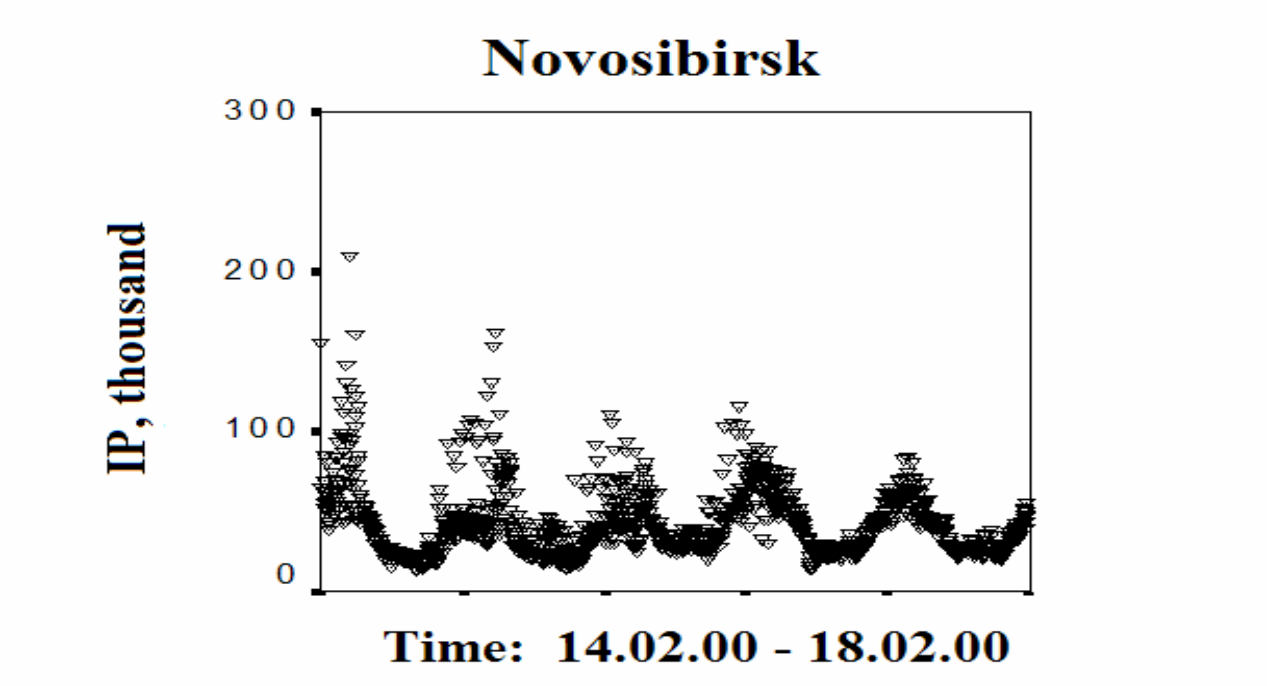
The typical traffic of Yekaterinburg at the RBNet-Teleglobe input till 5 days (5 min aggr.)



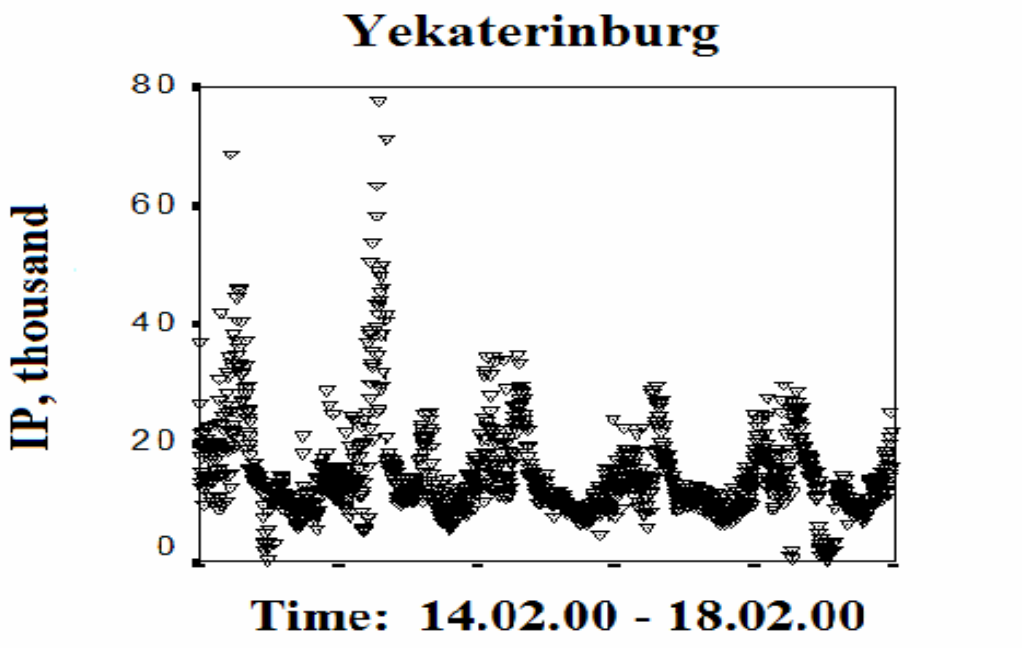
The number of the packets in the input traffic of RBNet-Teleglobe till 5 days (5 min aggr.)



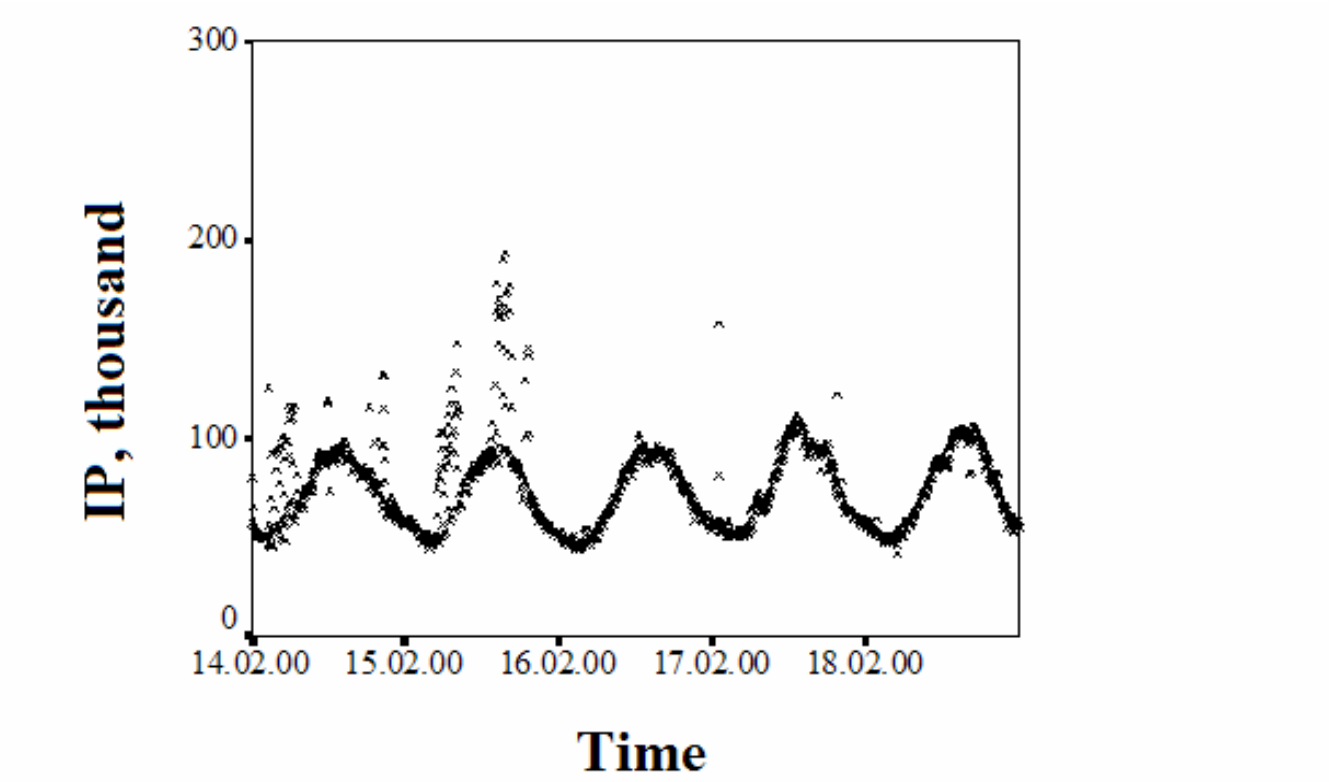
The number of the packets in the input traffic of Novosibirsk till 5 days (5 min aggr.)



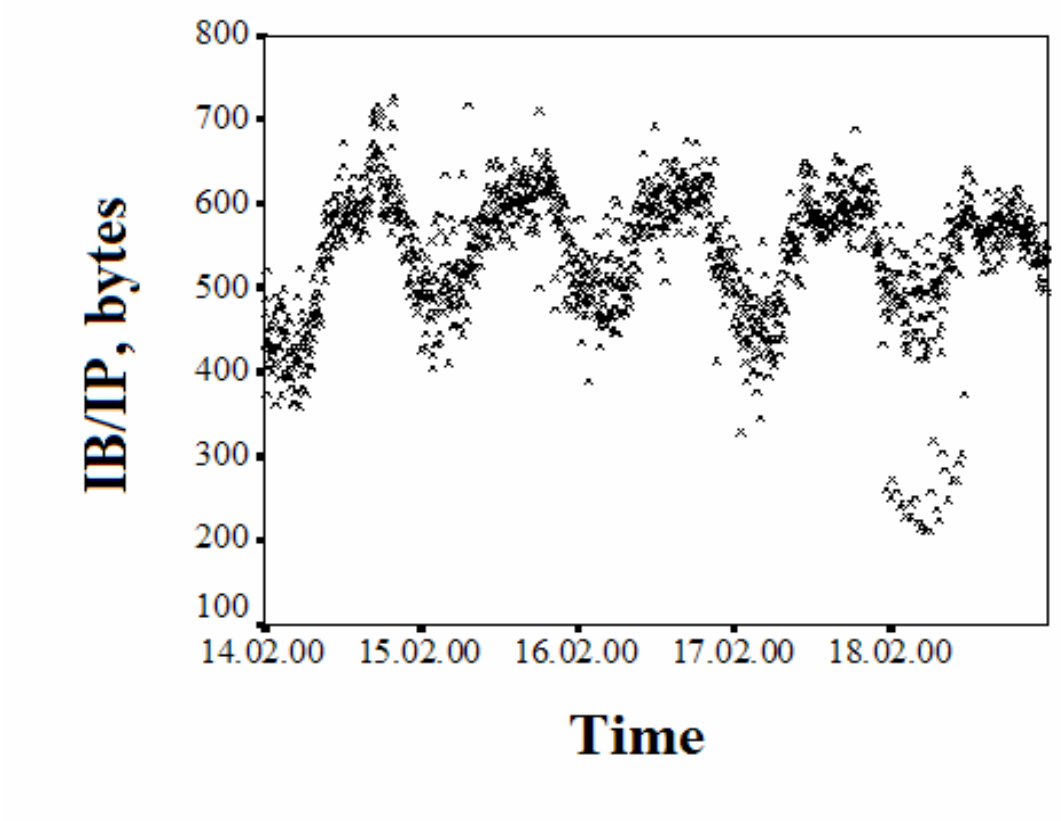
The number of the packets in the input traffic of Yekaterinburg till 5 days (5 min aggr.)



The number of the flows in the input traffic of RBNet-Teleglobe till 5 days (5 min aggr.)



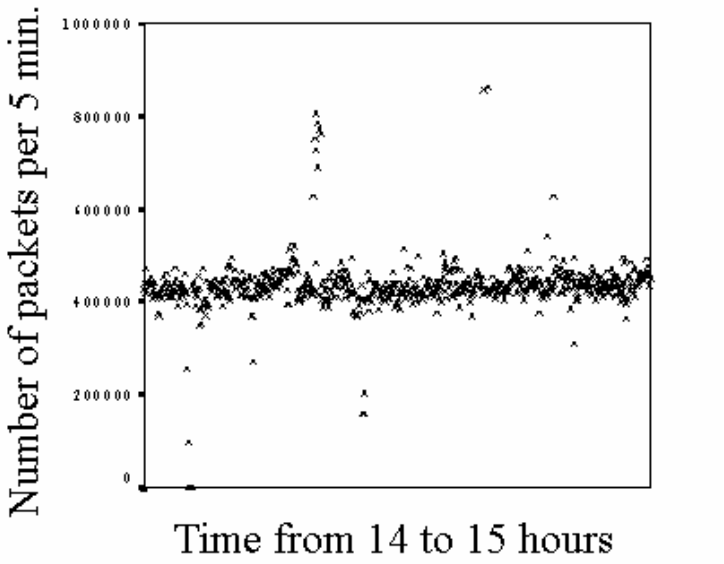
The average size of the packet in the input traffic of RBNet-Teleglobe till 5 days (5 min aggr.)

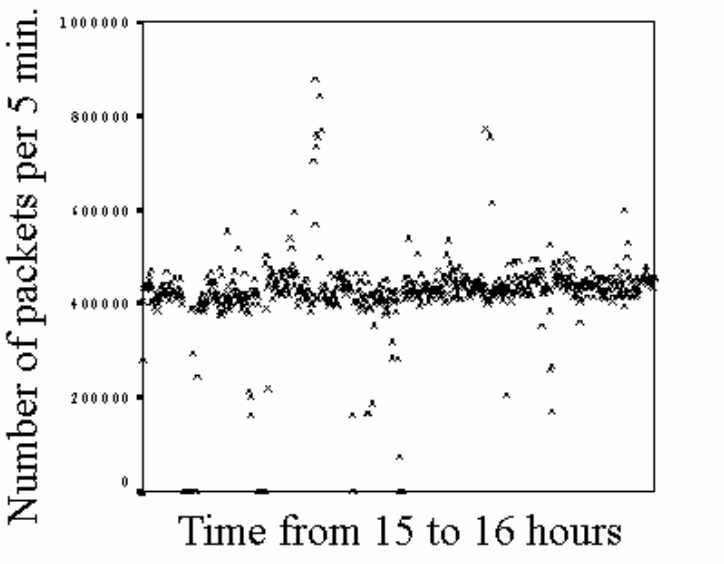


The specific features of traffic components distribution often prevent:

- the correct statistical inference;
- the estimations precision;
- the true determination of confidence levels and intervals;
- the estimations of correlation structure of data.

The number of packets
in the input traffic of RBNet-Teleglobe
in 14 and 15 hours
till the analysed time
(3.5 monthes, ~750 observations)



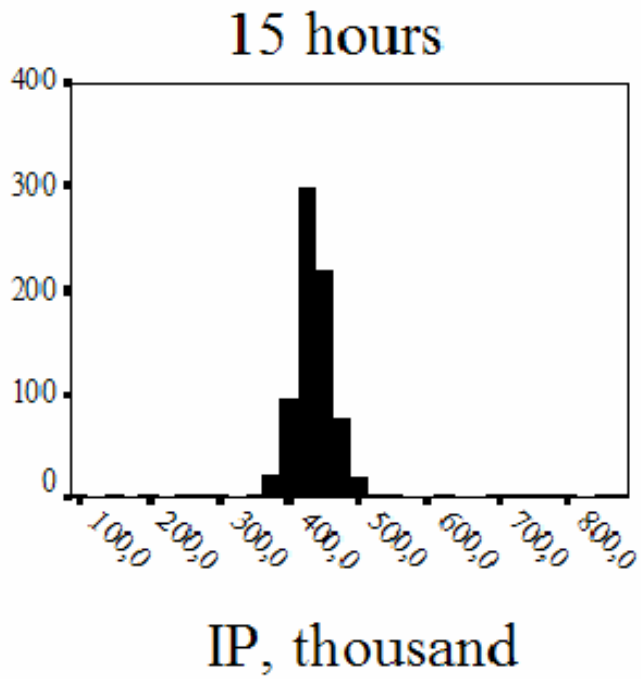


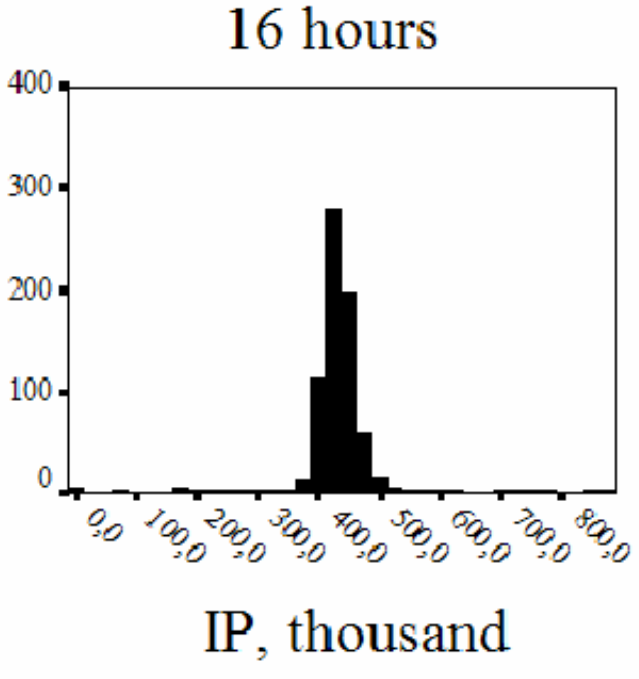
The comparison of the means and variances for the number of packets in the input traffic in 14-15 and 15-16 hours

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
IP	Equal variances assumed	6,702	,010	2,137	1474	,033
	Equal variances not assumed			2,126	1320,932	,034

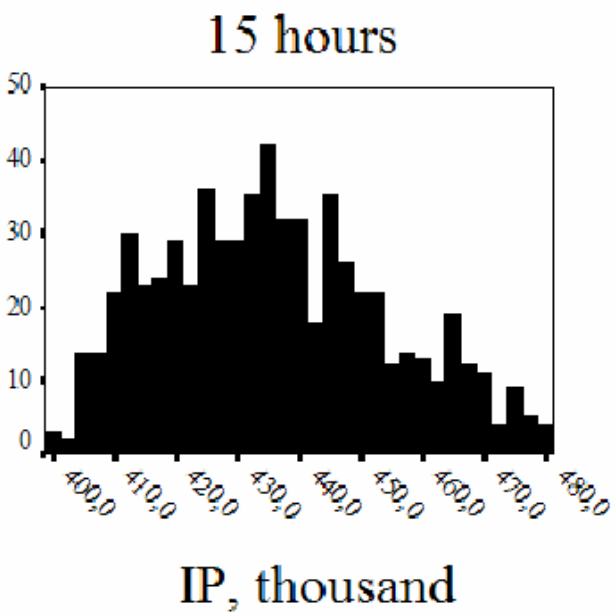
Histograms
of the not truncated data
(the number of the packets
per 5 min. in the input traffic)

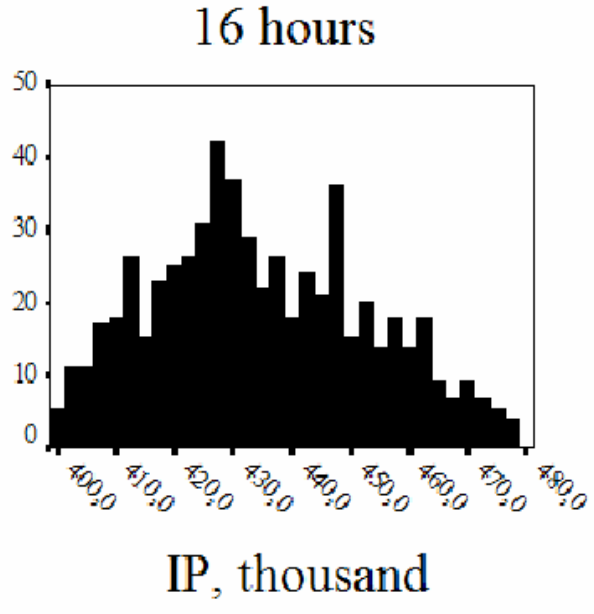




Histograms after truncation

(the number of the packets
per 5 min. in the input traffic)





The comparison of the means and variances for the number of packets after truncation in the input traffic in 14-15 and 15-16 hours

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
IP	Equal variances assumed	,128	,720	,963	1256	,336
	Equal variances not assumed			,963	1247,612	,336

The algorithm of the truncation based on goodness of fit Kolmogorov-Smirnov test for the conditional distributions

- 1. To compute the basic descriptive statistics:

  min, max, median, inter-quartile (IQ) range.
- 2. To choose a priory boundaries for the most probable values:

  median-1.5 IQ-range and median+1.5 IQ-range .
- 3. To fix the number of the truncation steps and to find the size of the step at the left and right.
- 4. To compute the two-sampling Kolmogorov-Smirnov statistics for the conditional distributions and its significance.
- 5. To stop truncations when the significance level of test is up the fix value (5%).

The computational protocol of the two-sampling goodness of fit Kolmogorov-Smirnov test for the conditional distributions of the truncated samples

hour	hour	boundaries		statistics	level	per cent of truncation		
		min	max		of significance	as a whole	left	right
2	6	1004	49797	1.421724	0.035	0	0	100
2	6	1103	49067	1.480169	0.025	0.6	0.2	99.6
2	6	1203	48337	1.459563	0.028	0.6	0.2	99.6
2	6	1302	47607	1.459563	0.028	0.6	0.2	99.6
2	6	1401	46878	1.459563	0.028	0.6	0.2	99.6
2	6	1501	46148	1.438906	0.032	0.6	0.2	99.6
2	6	1600	45418	1.438906	0.032	0.6	0.2	99.6
2	6	1699	44688	1.438399	0.032	0.7	0.2	99.5
2	6	1798	43958	1.396875	0.04	0.7	0.2	99.5
2	6	1898	43228	1.374317	0.046	0.7	0.2	99.5
2	6	1997	42498	1.374851	0.046	1	0.2	99.3
2	6	2096	41768	1.311171	0.064	1.1	0.2	99.1
2	6	2196	41039	1.290036	0.072	1.1	0.2	99.1
2	6	2295	40309	1.246839	0.089	1.2	0.2	99
2	6	2394	39579	1.225466	0.099	1.5	0.2	98.8
2	6	2494	38849	1.203031	0.111	1.8	0.2	98.4
2	6	2593	38119	1.136525	0.151	1.8	0.2	98.4
2	6	2692	37389	1.202037	0.111	2.2	0.2	98
2	6	2791	36659	1.158159	0.137	2.4	0.2	97.8
2	6	2891	35930	1.135787	0.151	2.7	0.2	97.6
2	6	2990	35200	1.135349	0.152	2.9	0.2	97.3
2	6	3089	34470	1.202026	0.111	3.3	0.2	96.9
2	6	3189	33740	1.065094	0.207	3.7	0.2	96.6
2	6	3288	33010	1.052745	0.218	4	0.2	96.2

The groups of the hours with the agreed conditional distributions of probabilities for the number of the packets in the input traffic for the various networks

Network/hour	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
RBNet-Teleglobe	1	1	1		2	2	2					
Novosibirsk	1	1	2	2	2	2	3	3				
Yekaterinburg		1	1	1	1	1						
Network/hour	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23 24**
RBNet-Teleglobe		3	3	4	4	4	4	5	5			
Novosibirsk	4	4	4	5	5	5	5		5			
Yekaterinburg	2	2	2	2				3	3			